Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 20 (2013) 385 – 390

Procedia Computer Science

Complex Adaptive Systems, Publication 3
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2013- Baltimore, MD

Complex Adaptive Systems Drive Innovations in Synthetic Biology

Melvin Greer[a], Manuel Rodriguez-Martinez[b], Jamie Seguel[b]*

[a]*Senior Fellow, Chief Strategist Lockheed Martin, 13560 Dulles Technology Dr, Herndon, VA 20171*
[b]*ECE Department, University of Puerto Rico, Mayaguez*

Abstract

The resultant outcome of synthetic biology research offers the promise of changing the way that we create energy, produce food, optimize industrial processing, and detect, prevent, and cure diseases. Researchers can now realize the significant benefits of the cloud computing model to reduce IT cost and complexity, enhance R&D collaboration, and improve bioinformatics to allow lifesciences organizations to focus on answering some of lifes most challenging questions.

Selection and peer-review under responsibility of Missouri University of Science and Technology

Keywords: IT; Cloud Computing; Big Data; Data Analytics; Life Science; Synthetic Biology; DNA; Bioinformatics; Complex Adaptive Systems

1. Introduction

Synthetic biology researchers require the best foundational tools, methodologies, and standards in order to assist them in questioning, uncovering, and exploring the hidden truths that will ultimately impact the human condition. These tools include a cellular chassis as an assembly and integration point of standardized biological parts, and a robust set of computational capabilities and techniques.

However, life sciences and pharmaceutical (Pharma) R&D organizations face significant challenges that are impacting their research endeavors.

- ✓ Produce more research at a lower cost
- ✓ Address their dramatic increase in the demand for computationally complex modeling and simulation
- ✓ Deal with shrinking budgets and limited resources that make it difficult to build the necessary foundational infrastructure to keep pace
- ✓ Improve the level of collaboration among fellow researchers in an increasingly distributed and globally dispersed R&D environment
- ✓ Contend with the ever increasing torrent of multi-source data
- ✓ Manage the distributed nature of research projects
- ✓ Improve the usability of computing resources to non-IT savvy researchers

Life science is any branch of natural science, such as biology, that deals with living organisms, their structure, life processes, organization and relationships.

* Corresponding author.
E-mail address: jaime.seguel@upr.edu

Selection and peer-review under responsibility of Missouri University of Science and Technology
doi:10.1016/j.procs.2013.09.291

Synthetic biology is defined as "the design and construction of new biological parts, devices and systems and the redesign of existing natural biological systems for useful purposes"[1]. It is an emerging life science field that combines science and technology with the aim of applying engineering principles in the rational design of biological systems. Synthetic biology strives to make the engineering of biology simpler and more predictable.

Synthetic biology itself can be divided into two categories: advancing basic knowledge through research activities and its practical application in the creation of new products.

Synthetic biology, sometimes referred to as *synbio*, is an offshoot of traditional genetic engineering. Instead of randomly tweaking the genetic blueprints of a desirable trait within a living organism, the goal is to be able to effectively engineer organisms, and assign them novel functions that will lead to potential applications in different areas.

Complex Adaptive Systems (CAS) are diverse systems that are comprised of multiple, interconnected elements and are adaptive in their capacity to respond and evolve as they interact with their surroundings. The human body and life itself can be considered to be examples of Complex Adaptive Systems. CAS addresses many of the issues, such as adaptability, scalability, self-organization, reliability, robustness, and self-repair, that are required in the development of ever-sophisticated social or natural systems. Advances in synthetic biology will have to meet some larger challenges that derive from the nature of Complex Adaptive Systems.

2. The Synthetic Biology Process

The modification of the genetic characteristics of plants and animals in order to manage certain traits or characteristics has been practiced for thousands of years. A gene is the basic physical and functional unit of heredity that is capable of transmitting characteristics from one generation to the next. The discovery of DeoxyriboNucleic Acid (DNA) in 1953 was a significant breakthrough due to its importance as the carrier of heritable genetic information

Research efforts such as the Human Genome Project (HGP) whose primary goal was determining the sequence of chemical base pairs which make up DNA, and of identifying and mapping the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint, was considered another important milestone.

In May 2010, a team of scientists led by J. Craig Venter became the first to successfully create what some have described as "synthetic life". This was performed by synthesizing commercially made strands of DNA containing an entire bacterium genome, and introducing it into another cell.

Initially gene manipulation was painstakingly slow, subject to simpler activities, and employed a *trial and error* approach to reach a satisfactory result. In order to overcome these limitations, newer approaches were developed which combine engineering and biological techniques in order to manipulate DNA. Advances in Information Technology (IT)-related technologies ranging from stand-alone laboratory-based equipment, PC-based software products, on-premise computational and data storage, and cloud computing capabilities, have been contributing factors in moving biological engineering activities forward. Through the use of these new engineering methodologies and foundational tools, synthetic biology researchers can study, alter, create, and re-create highly complex DNA sequences, genes, and natural biological systems.

There is a distinction between the work of a genetic engineer and a synthetic biology engineer. What separates synthetic biology from genetic engineering is that rather than altering an already existent DNA strand, synthetic biology puts these genetic components together from scratch to build an entirely new strand of DNA which is then placed into an empty living cell. These new cells offer functions that are not found in nature.

The application of engineering to biology applies the traditional *Analysis - Design - Build - Test'* engineering cycle. Also following the engineering paradigm, standard parts or components are used for construction of new biosystems. The basic components that a synthetic biology researcher works with are standard biological parts or "building blocks" such as genes, proteins, and DNA sequences that can be assembled to "program" living cells and control an organism's functions.

To facilitate the standardization and assembly of these parts, synthetic biologists have begun to adopt a "Biobrick™" standard. Akin to traditional "engineered" components, their performance, behavior, and parameters are characterized in an associated descriptive datasheet. Through the use of this approach, new Biobrick™ parts, as depicted in Figure 1, can be designed and produced, assembled into construction intermediates and ultimately into useful biological devices. .

Figure 1 – A representative BioBrick™ part depiction

Biobricks™ are already widely available from commercial websites. The BioBricks™ Foundation [2] is charting a technical standards framework that will serve as the driver and promoter of a high-quality, technical-standards process for synthetic biology based on these parts.

Engineering new forms of life starts with setting up a biological assembly line. This process resembles computer software development process in that scientists assemble blocks of genetic "code" into instructions for tiny cellular machines, such as DNA. BioBrick™ parts are composable; allowing endless numbers of these parts to be pieced together to form complex systems.

The engineering challenge of synthetic biology lies in the timely design of genetic networks that will rewire cells to produce the expected molecules. In order to achieve breakthroughs in life sciences discovery, there is a vital need carry out modeling and simulations, perform data gathering and analytics at a large scale, manage data from instruments, and share and visualize the results with scientific peers on a global basis.

Cloud computing and Big Data offer a significant positive impact on synthetic biology research effectiveness, efficiency, and timeliness.

.

3. The Impacts of Cloud Computing

Cloud computing products and services can be broken down into the following categories.

- Infrastructure-as-a-Service (IaaS) allows end users to fully outsource the provisioning of servers, software, and/or networking.
- Platform-as-a-Service (PaaS) offers a hosted computing platform that allows users to deploy applications without having to buy and manage the required hardware and underlying software layers.
- Software-as-a-Service (SaaS) refers to software licensed by a provider to users on a contractual or utility basis.
- Storage as a Service (StaaS) - Although not defined as a traditional cloud computing delivery model, it offers cloud provider based digital storage

Computational Resources On Demand

Advanced computational problems that are essential in synthetic biology, such as DNA sequencing, fabrication of genes, modeling how synthetic genes behave, and precisely measuring gene behavior require complex computing capabilities that have traditionally not been readily available to researchers.

The cloud allows researchers to focus their time and attention on the complex problems they want to solve by removing the burden for them to be IT experts. The ability to rapidly provision resources is critical to provide elastically in scaling to thousands of cores in response to a researcher's needs. Through the use of Cloud Service Providers (CSP), a researcher can gain use of a flexible infrastructure on demand ordered up by a credit card and available immediately. Furthermore, the cloud can be switched on or off as required, and researchers can easily go from using single servers to large clusters and back again and release them back to the cloud when no longer needed. Researchers need something simple and reliable with minimal IT support that is available today in a variety of cloud service delivery and deployment models.

A number of life sciences-oriented companies provide "middleware" and cloud services such as SaaS that facilitate next-generation DNA sequencing, protein docking, modeling and simulation, and data mining.

Much of the interest and early activity in cloud computing for life sciences focuses on DNA sequencing. . Next-Generation Sequencing (NGS), a fundamentally different approach to sequencing, has triggered numerous groundbreaking discoveries and ignited a revolution in genomic science. The demands for NGS data processing

suffer large peaks and valleys, and rapid growth in utilization with the staggering quantities of data generated. Some instances of NGS systems generate qualitatively different kinds of data, which will present still new informatics challenges. This represents a quantum leap in the amount of data generated and the levels of infrastructure needed to keep pace with researchers' expectations

Collaboration Resources

The importance of collaborations that facilitate the sharing of knowledge necessary to realize the full research potential, both inter and intra industry and academia, is clear.

Collaboration can take many forms and can be with research teams, the larger research community, academic institutions, vendors, customers, or regulatory bodies. Industry consolidation and globalization have shifted workflows within the research communities that now require scientists from around the world to collaboratively work on joint projects. These research projects and processes may be segmented into components that are distributed to the most effective collaborators and facilities.

Traditional IT solutions would require complex security integration to facilitate end user access to the Internet for research data or applications. Cloud computing tools and services can help to meet research challenges that facilitate productivity and communications across multiple time zones and through multiple end user computing devices.

The latest research data and tools can reside in centralized secure cloud-based common access workspace repositories allowing scientists to access documents and update their research regardless of their physical work location. In order to facilitate effective and scalable collaborative communities, cost-effective cloud based email, social media, video chat, shared whiteboard markup tools, and full video teleconferencing can help bring researchers together to solve complex problems.

While cloud computing is opening up a wide range of opportunities for innovation for researchers to share data across the world, it must be noted that these opportunities open up new areas of risk and maintaining security and privacy of data is paramount.

Crowd Sourcing

The Internet has opened up the potential of tapping into the collective intelligence of the public for mass collaboration with the general public. This process, known as CrowdSourcing, can provide a new approach for large groups to share thoughts, collaborate, and build off of one another's ideas. While not truly a cloud computing technology, it leverages Internet and cloud computing capabilities.

Foldit, which is considered a prime example of CrowdSourcing, has led to key scientific discoveries. For example, players--many of whom have no formal scientific training--came up with the basic structure of an enzyme used to study AIDS after researchers had fallen short in doing so for more than a decade.

Researchers utilized the CrowdSourcing paradigm to turn to software programmers scattered around the world to help solve a computational biology problem. The resultant solution was an unqualified success as it provided an improved algorithmic solution to the deeply biological problem of analyzing the makeup of genes that produce proteins involved in the immune system's ability to identify microbes [3].

Data Storage and Archival

The lifeblood of life sciences organizations is the massive amounts of data stored and referenced during vital R&D activities. However, the IT infrastructure is crumbling under the ever-increasing torrent of genomic and related scientific data.

An example of the data deluge was noted in a February 2011 Science magazine article titled "Will Computers Crash Genomics?" Article author Elizabeth Pennisi, stated that. "a single DNA sequencer can now generate in a day what it took 10 years to collect for the Human Genome Project," [4]. She further noted that sequencing centers have produced data sets so large they have to be mailed physically on disks and drives due to the significant amount of time and bandwidth considerations to transfer them electronically over the Internet. For instance, a sequencing center can generates one chain of genomic data equaling 9 TB in a single week. However, data storage requirements for next-generation sequencers have been alleviated to some extent by increasingly processing mounds of raw image data in real time and providing users with smaller quantities of processed sequence data to feed their analysis pipelines.

In the pre-cloud computing era, this type of data would typically be stored in secure on-premise systems with costly operational overhead. In many cases valuable data would either deleted based on operational costs or lack of enough available storage capacity. Depending on the type of data stored, there may be legal requirements to archive data for years that drastically raises the cost of storage.

The Storage as a Service business model provides a good alternative for research organizations that lack the capital budget and/or technical personnel to implement and maintain their own storage infrastructure.

The Impact of Big Data

In order to perform their research, scientists are increasingly having to deal with multi-structured data drawn from Internet connected devices, open source data, digital records and documents, study databases, and spreadsheets.

In the last five years, more scientific data has been generated than in the entire history of mankind," according to Winston Hide, associate professor of bioinformatics at Harvard School of Public Health. "You can imagine what's going to happen in the next five" [5]. And this data goes beyond simple linear data to include highly dimensional data, which is fundamentally different in scale. The ability to generate data far outstrips the ability to analyze it.

Big Data is a general term used to describe the velocity, variety and volumes of unstructured and semi-structured data an organization creates. From a researcher's perspective, the primary goal for looking at Big Data is to discovery of data patterns. Big Data analytics is often associated with cloud computing because the analysis of large data sets in real-time requires a framework (Hadoop, NoSQL, MapReduce) to distribute the work among hundreds to thousands of computers.

Hadoop can help solve the challenge of big life sciences datasets as it supports parallelization, offers good composability, and can map problems, such as genomics, through Map Reduce. This new improved Big Data cloud-based infrastructure and new computational tools to render the data suitable for meaningful analysis, allow it to maintain the massive torrent of data that keeps increasing from new and upcoming research.

The Impact of Data Analytics and Search

There is a wealth of data analytic tools and techniques for life science researchers to use upon the exploding data collections. The next-generation sequencing already has bioinformaticians having a difficult time keeping up with the needs of biologists to convert that data to useful information.

Bioinformatics is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data. A major goal of bioinformatics is the development of computational algorithms, image processing and software tools to generate useful biological knowledge. Bioinformatics uses biological information and computational methods to discover knowledge. This can range from the selection and characterization of components, to development of the design rules, and the creation of testing protocols. It additionally aids in sequencing and annotating genomes and their observed mutations. Large-scale data analysis is now gaining equal importance with theory, experimentation and simulation.

Search tools provide an increasingly critical capability for the life science researcher. At a minimum, good search tools allow search of diverse information repositories in order to easily find and leverage the work already conducted by fellow researchers around the world. Data such as demographics, public health records, etc can facilitate research activities, which are increasingly being maintained in cloud-based environments that may be obtained by through search tools or standards-based APIs.

The availability of the voluminous amounts of whole genome sequences has opened up new avenues for genome-wide innovative genome design. New tools for analyzing these sequences and mining the genome design are being developed. Through the analyses of the complete genomes for similarities/differences and structural features, it has become possible to identify new genetic targets.

Visualization tools used to provide the mechanism to spot the trends in massive amounts of scientific data. New machine learning algorithms approaches are superseding visualization in data collection that are far too large or complex to visualize.

4. Summary

Synthetic biology joins biology together with engineering to improve the design and optimization of cellular behavior. By building biological systems from component parts, this emerging field is expected to produce proteins and molecules with specifically engineered functions. The many positive benefits and power of the cloud is already unleashing new research tools, design processes and methodologies to support synthetic biology. As synthetic

biology matures into a robust engineering discipline, it should be capable of transforming the biotechnology, pharmaceutical, healthcare, and chemical industries.

References

[1] http://syntheticbiology.org
[2] http://biobricks.org/
[3]http://www.bostonglobe.com/business/2013/02/11/crowdsourcing-innovation-harvard-study-suggests-prizes-can-spur-scientific-problem-solving/JxDkOkuIKboRjWAoJpM0OK/story.html
[4]http://www.sciencemag.org/content/331/6018/666.short
[5]http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs
[6] https://cloud.google.com/products/compute-engine
[7] http://www.helix-nebula.eu/
[8]: http://www.numatix.net/
[9] http://www.pacificbiosciences.com
[10] http://www.genomics.cn/en
[11]: http://www.numerate.com/
[12] http://www.genomequest.com/
[13] http://www.geospiza.com/
[14] http://www.nervanix.com/
[15] http://aws.amazon.com/glacier/
[16] http://www.eaglegenomics.com/
[17] http://aws.amazon.com/articles/2272
[18] http://www.thefreedictionary.com/**bioethics**
[19]http://www.bioethics.gov/documents/Executive-Order-Establishing-the-Bioethics-Commission-11.24.09.pdf
[20]http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-wagner-1-7-2011.pdf - Presidential Commission for the Study of Bioethical Issues, New Directions: The Ethics of Synthetic Biology and Emerging Technologies (Washington, D.C.: Government Printing Office, 2010
[21]http://www.synbioproject.org/process/assets/files/6420/final_synbio_funding_web2.pdf
[22]http://www.prweb.com/releases/synthetic_biology/genetic_engineering/prweb4247114.htm
[23]http://www.onr.navy.mil/en/Media-Center/Fact-Sheets/Synthetic-Biology.aspx